San José State University

Math 261A: Regression Theory & Methods

Multicollinearity (and Model Validation)

Dr. Guangliang Chen

Multicollinearity (and Model Validation)

This lecture is based on the following part of the textbook:

- Multicollinearity: Sections 3.10, 9.1 – 9.5
- Model validation: Sections 11.1–11.2

Outline of the presentation:

- Effects of multicollinearity
- How to detect multicollinearity
- How to deal with multicollinearity
- Ridge regression (and model validation)

Introduction

Recall the multiple linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \tag{1}$$

Given a set of n observations

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

the least-squares (LS) estimator of $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

provided that $\mathbf{X}'\mathbf{X}$ is nonsingular (or invertible).

Remark. The condition on $\mathbf{X}'\mathbf{X}$ holds true if and only if $\mathbf{X}$ is of full column rank, i.e, the columns of $\mathbf{X}$ are linearly independent.

A serious issue in multiple linear regression is **multicollinearity**, or near-linear dependence among the regression variables, e.g., $x_3 = 2x_1 + 3x_2$.

- $\mathbf{X}$ won't be of full rank (and correspondingly, $\mathbf{X}'\mathbf{X}$ is not invertible)
- Redundant predictors carry no new information about the response:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 = \hat{\beta}_0 + (\hat{\beta}_1 + 2\hat{\beta}_3)x_1 + (\hat{\beta}_2 + 3\hat{\beta}_3)x_2$$

- The estimated slopes in the regression model will be arbitrary

$$y = x_1 + x_2 + 2x_3 = 3x_1 + 4x_2 + x_3 = 5x_1 + 7x_2 = \cdots$$

The multicollinearity issue may arise for the following reasons:

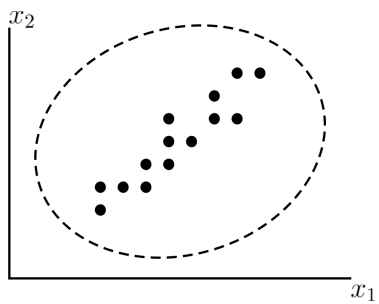
- **Too many predictors collected** (without noticing their correlation), e.g.,

$$x_1 = \text{arm length}, \quad x_2 = \text{leg length}, \quad x_3 = \text{height}, \ \ldots$$

- **Higher-order terms** used in the model (e.g. polynomial regression):

$$x_1 = x, \quad x_2 = x^2, \quad x_3 = x^3, \ \ldots$$

- **Poor choice of data collection**



This lecture presents the following in detail:

- **Effects** of multicollinearity on a multiple regression model
- **Tools** to determine whether a model has multicollinearity problems
- **Ways** to deal with multicollinearity problems in a model

Effects of multicollinearity

To study the effects of multicollinearity on regression, we consider a multiple linear regression problem with two predictors x_1, x_2:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \quad i = 1, \ldots, n$$

where

- The response y is centered ($\bar{y} = 0$), and
- Both predictors have been normalized to have unit length, i.e.,

$$\bar{x}_1 = \bar{x}_2 = 0, \qquad \text{and} \quad S_{11} = S_{22} = 1$$

such that $\mathbf{X}'\mathbf{X}$ **is the correlation matrix between them**.

Let the correlation between x_1, x_2 be

$$r_{12} = \frac{\sum(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{\sqrt{\sum(x_{i1} - \bar{x}_1)^2 \sum(x_{i2} - \bar{x}_2)^2}} \quad \longrightarrow \quad \mathbf{X}'\mathbf{X} = \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}$$

It follows that

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 \cdot \frac{1}{1 - r_{12}^2} \begin{pmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{pmatrix}$$

That is,

$$\mathrm{Var}(\hat{\beta}_1) = \frac{1}{1 - r_{12}^2}\sigma^2 = \mathrm{Var}(\hat{\beta}_2), \quad \mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) = -\frac{r_{12}}{1 - r_{12}^2}\sigma^2.$$

Clearly,

$$\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}(\hat{\beta}_2) \to \infty \quad \text{as } r_{12} \to 1 \text{ (or } -1).$$

This means that in the case of strong multicollinearity ($|r_{12}| \approx 1$), a slightly different sample can lead to vastly different estimates of the model parameters.

Another interesting phenomenon is that **if there is multicollinearity in a model, the parameters tend to be overestimated in magnitude**.

To see this, consider the squared distance between the estimator $\hat{\boldsymbol{\beta}}$ and the true value $\boldsymbol{\beta}$:

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 = \sum_j (\hat{\beta}_j - \beta_j)^2$$

Taking expectation gives that

$$\begin{aligned}
\mathrm{E}\left(\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|^2\right) &= \sum_j \underbrace{\mathrm{E}\left[(\hat{\beta}_j-\beta_j)^2\right]}_{=\mathrm{Var}(\hat{\beta}_j)} \\
&= \mathrm{trace}\left(\mathrm{Var}(\hat{\boldsymbol{\beta}})\right) \\
&= \sigma^2 \cdot \mathrm{trace}\left((\mathbf{X}'\mathbf{X})^{-1}\right) \\
&= \sigma^2 \cdot \frac{2}{1-r_{12}^2}
\end{aligned}$$

Thus, in the case of strong multicollinearity, $\hat{\boldsymbol{\beta}}$ is (on average) far from $\boldsymbol{\beta}$.

On the other hand,

$$\begin{aligned}\mathrm{E}\left(\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|^2\right) &= \mathrm{E}\left(\|\hat{\boldsymbol{\beta}}\|^2+\|\boldsymbol{\beta}\|^2-2\boldsymbol{\beta}'\hat{\boldsymbol{\beta}}\right)\\ &= \mathrm{E}\left(\|\hat{\boldsymbol{\beta}}\|^2\right)+\|\boldsymbol{\beta}\|^2-2\boldsymbol{\beta}'\mathrm{E}\left(\hat{\boldsymbol{\beta}}\right)\\ &= \mathrm{E}\left(\|\hat{\boldsymbol{\beta}}\|^2\right)-\|\boldsymbol{\beta}\|^2\end{aligned}$$

Therefore,

$$\begin{aligned}\mathrm{E}\left(\|\hat{\boldsymbol{\beta}}\|^2\right) &= \|\boldsymbol{\beta}\|^2+\mathrm{E}\left(\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|^2\right)\\ &= \|\boldsymbol{\beta}\|^2+\frac{2\sigma^2}{1-r_{12}^2}\end{aligned}$$

This indicates that in the case of strong multicollinearity, the norm (length) of the vector $\hat{\boldsymbol{\beta}}$ is (on average) much larger than that of $\boldsymbol{\beta}$.

Detecting multicollinearity

Ideally, we would like to know not only whether there is multicollinearity in the model, but also what degree of problem we have (weak, moderate, strong, etc.) and determine which predictor variable(s) cause the problem.

1. **Scatterplot/correlation matrix**: This is a good first step but can only reveal near-linear dependence between a pair of predictors.
2. **Variance inflation factors (VIFs)**: Can detect near-linear dependence among any number of predictors.
3. **Condition number of the correlation matrix** ($\kappa = \lambda_{\max}/\lambda_{\min}$): A large value ($> 1000$) indicates strong multicollinearity in the data.

Detecting correlation between two predictors

When there is a clear linear dependence between two predictors, this can be detected by

- looking at the scatter plot matrix of all predictors ⟵ can be subjective
- computing the pairwise correlation scores ⟵ better

We demonstrate this with the *Longley's Economic Regression Data*.[1] [2]

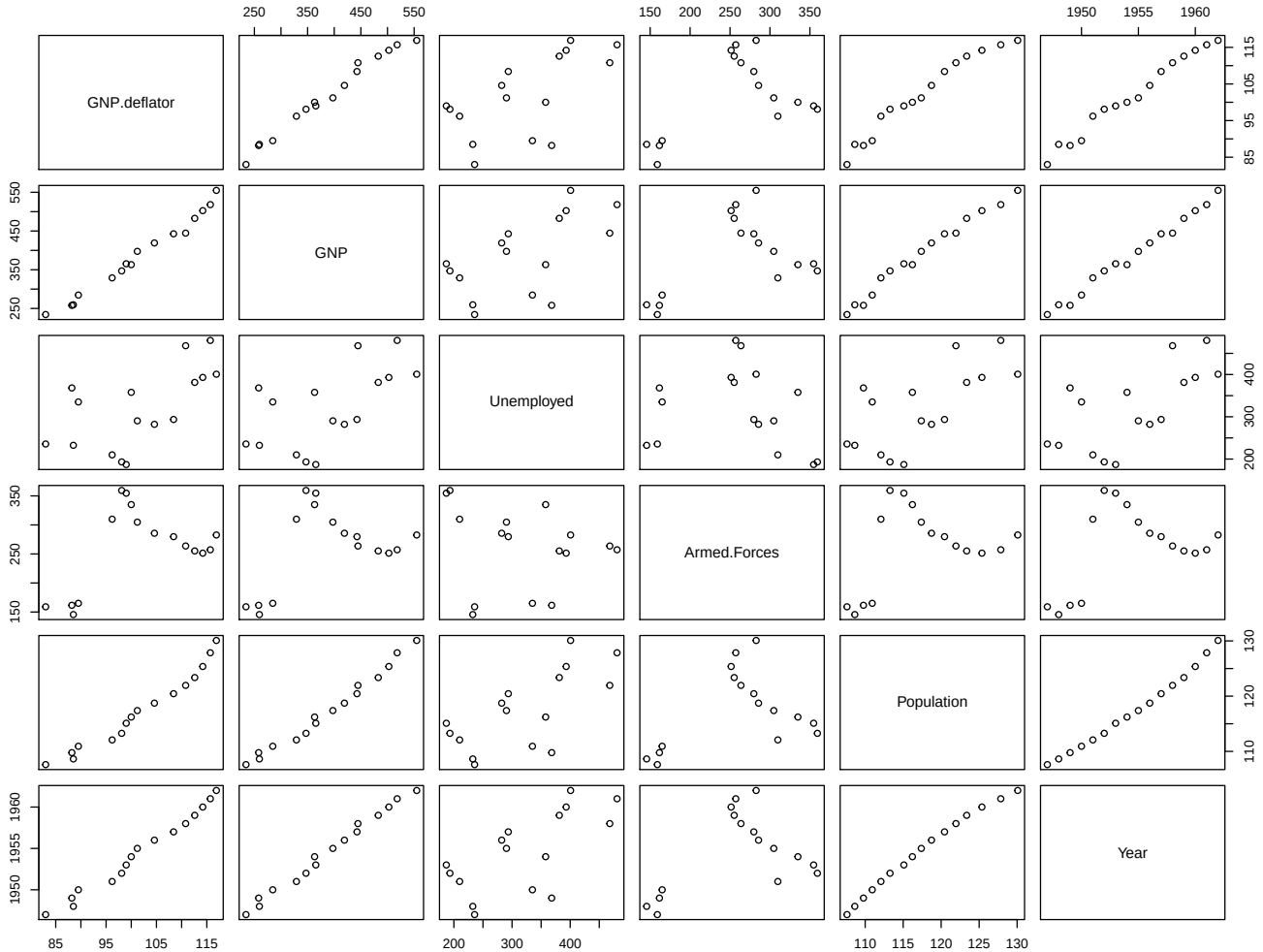
[1] `https://rweb.webapps.cla.umn.edu/R/library/datasets/html/longley.html`
[2] `https://www2.stat.duke.edu/courses/Fall16/sta721/slides/Ridge/ridge.html`

Example: Longley's Economic Regression Data

There are 7 economical variables, observed yearly from 1947 to 1962 (n=16):

1. GNP.deflator: GNP implicit price deflator (1954=100)
2. GNP: Gross National Product.
3. Unemployed: number of unemployed.
4. Armed.Forces: number of people in the armed forces.
5. Population: 'noninstitutionalized' population $\geq$ 14 years of age.
6. Year: the year (time).
7. Employed (response): number of people employed.

Multicollinearity (and Model Validation)



Matrix of pairwise correlations

	GNP.deflator	GNP	Unemployed	Armed.Forces	Population	Year
[1,]	1.000	0.992	0.621	0.465	0.979	0.991
[2,]	0.992	1.000	0.604	0.446	0.991	0.995
[3,]	0.621	0.604	1.000	-0.177	0.687	0.668
[4,]	0.465	0.446	-0.177	1.000	0.364	0.417
[5,]	0.979	0.991	0.687	0.364	1.000	0.994
[6,]	0.991	0.995	0.668	0.417	0.994	1.000

Conclusion: The following regressor pairs are all highly correlated:

(1,2), (1,5), (1,6), (2,5), (2,6), (5,6).

Detecting correlation among three or more predictors

To check for multicollinearity among any number (k) of predictors, we regress each single predictor $x_j,\ j = 1, \ldots, k$ on the remaining ones, i.e.,

$$x_j \ \sim \ x_1 + \cdots + x_{j-1} + x_{j+1} + \cdots + x_k$$

and compute the corresponding coefficients of determination R_j^2.

A large value of R_j^2 indicates strong linear dependence of x_j on the other regressors, thus implying multicollinearity of the predictors in the model.

However, the above process requires fitting k separate models. We present a shortcut approach next.

Given a set of observations $\mathbf{X}$ from a multiple linear regression model, let $\mathbf{W}$ be the rescaled data matrix corresponding to unit length scaling:

$$\mathbf{X} = [1\ \mathbf{x}_1\ \ldots \mathbf{x}_k] \qquad \longrightarrow \qquad \mathbf{W} = [\mathbf{w}_1\ \ldots \mathbf{w}_k]$$

That is,

$$w_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{S_{jj}}}, \quad i = 1, \ldots, n,\ j = 1, \ldots, k$$

Note that $\mathbf{W}'\mathbf{W}$ is the correlation matrix for the predictors in the model.

The main diagonal elements of its inverse, $(\mathbf{W}'\mathbf{W})^{-1}$, are called the **variance inflation factors (VIF)** of the regressors.

Theorem 0.1. For each regressor $j = 1, \ldots, k$,

$$R_j^2 = 1 - \frac{1}{\mathrm{VIF}_j} \qquad \left(\text{or } \mathrm{VIF}_j = \frac{1}{1 - R_j^2} \right)$$

Remark. Consider the following cases:

- When x_j is nearly a linear combination of the other regressors:

$$R_j^2 \approx 1 \qquad \longrightarrow \qquad \mathrm{VIF}_j \text{ is large}$$

- When x_j is orthogonal to all the other regressors:

$$R_j^2 = 0 \qquad \longrightarrow \qquad \mathrm{VIF}_j = 1$$

How to use VIFs to detect multicollinearity:

- The larger these factors are, the more you should worry about multicollinearity in your model.

  A rule of thumb is that if for some j,

  $$\text{VIF}_j > 10$$

  then multicollinearity is high among the predictors.

- On the other extreme, all VIFs being 1 mean that the predictors are orthogonal to each other.

Example: Longley's Economic Regression Data (cont'd)

	GNP.deflator	GNP	Unemployed	Armed.Forces	Population	Year
[1,]	1.000	0.992	0.621	0.465	0.979	0.991
[2,]	0.992	1.000	0.604	0.446	0.991	0.995
[3,]	0.621	0.604	1.000	-0.177	0.687	0.668
[4,]	0.465	0.446	-0.177	1.000	0.364	0.417
[5,]	0.979	0.991	0.687	0.364	1.000	0.994
[6,]	0.991	0.995	0.668	0.417	0.994	1.000
VIF	135.5	1788.5	33.6	3.6	399.2	759.0

Condition number: $\kappa(\mathbf{X}'\mathbf{X}) = 33,076,481,\ \kappa(\mathbf{W}'\mathbf{W}) = 8,908,139$

Another way of detecting the multicollinearity among all the predictors in the data set is through examining the condition number of $\mathbf{X}\mathbf{X}'$.

Def 0.1. The **condition number** of a square matrix $\mathbf{A}$ is defined as the ratio of its largest eigenvalue to its smallest eigenvalue:

$$\kappa(\mathbf{A}) = \lambda_{\max}/\lambda_{\min}$$

Observations:

- A singular matrix has a condition number of infinity (worst).
- A nearly singular matrix has a "large" condition number.
- The identity matrix (or any constant multiple of it) has a condition number of 1 (smallest possible).

Remark. Generally,

- If the condition number is less than 100, there is no serious problem with multicollinearity.
- Condition numbers between 100 and 1000 imply moderate to strong multicollinearity.
- Condition numbers bigger than 1000 indicate severe multicollinearity.

Remark. Scaling the predictors helps reduce the condition number!

- We can compute the condition number of $\mathbf{X}\mathbf{X}'$ (where $\mathbf{X}$ represents the original design matrix). $\longleftarrow$ The fitted model is $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$
- We could choose to scale the predictors first and then examine the condition number of the correlation matrix $\mathbf{W}\mathbf{W}'$. $\longleftarrow$ The fitted model is $\hat{\mathbf{y}} = \mathbf{W}\hat{\mathbf{b}}$
- The condition number of the correlation matrix $\mathbf{W}\mathbf{W}'$ indicates the true extent of multicollinearity in the data.

(See another R demonstration with the National Football League dataset - Table B1 in the Appendix of the textbook)

Dealing with multicollinearity

1. If the multicollinearity is caused through bad sampling in predictor space, then **adding more data values** at appropriate points can correct the problem.

2. If the choice of the model causes the multicollinearity, then it may be possible to **reformulate the model** to fix the problem (e.g., by centering the variables in quadratic regression problems).

3. **Ridge regression** (and LASSO). ⟵ Will be covered next

4. Use **model selection** methods to eliminate redundant predictors from the model. ⟵ To be covered in Chapter 10

Ridge regression

We have seen that for least-squares regression, the magnitude of $\hat{\beta}$ is inflated if the data contains multicollinearity. That means that confidence intervals for the slope parameters will tend to be wide and estimation of the slopes can be unstable.

Ridge regression (for fitting a multiple regression model $y \sim x_1 + \cdots + x_K$) is like least squares regression but **shrinks the estimated coefficients towards zero** to fix the magnitude inflation.

To do this, Ridge regression assumes that the model has no intercept term, or both the response and the predictors have been centered so that $\hat{\beta}_0 = 0$.

It then fits a linear model by penalizing the coefficient vector $\hat{\boldsymbol{\beta}}$ for having a large magnitude ($\|\hat{\boldsymbol{\beta}}\|^2 = \sum_{j=1}^{k} \hat{\beta}_j^2$):

$$\min_{\hat{\boldsymbol{\beta}}} \underbrace{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2}_{\text{fitting error}} + \lambda \underbrace{\|\hat{\boldsymbol{\beta}}\|^2}_{\text{penalty}} \qquad \longrightarrow \hat{\boldsymbol{\beta}}_R \quad \text{(Ridge estimator)}$$

Here, $\lambda \geq 0$ is a tradeoff parameter (amount of shrinkage), which controls the strength of the penalty term:

- When $\lambda = 0$, we get the least squares estimator: $\hat{\boldsymbol{\beta}}_R = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$
- When $\lambda = \infty$, we get $\hat{\boldsymbol{\beta}}_R = \mathbf{0}$
- Increasing the value of λ forces the norm of $\hat{\boldsymbol{\beta}}$ to decrease, yielding smaller coefficient estimates (in magnitude).

Remark.

- For a finite, positive value of λ, we are balancing two tasks: fitting a linear model and shrinking the coefficients.
- Choosing an appropriate value of λ is important, yet not easy. We will address it later.
- The penalty term $\|\hat{\boldsymbol{\beta}}\|^2$ would be unfair to the different predictors, if they are not on the same scale. Therefore, if we know that the variables are not measured in the same unit, we typically first perform unit normal scaling on the columns of $\mathbf{X}$ (to standardize the predictors), and then perform ridge regression.

The solution to the Ridge regression problem always exists and is unique, even when the data contains multicollinearity.

Theorem 0.2. The ridge estimator $\hat{\boldsymbol{\beta}}_R$ is given by the following

$$\underbrace{(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})}_{\text{PD, invertible}}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \quad \longrightarrow \quad \hat{\boldsymbol{\beta}}_R = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$$

Proof. Differentiate the Ridge objective function with respect to $\hat{\boldsymbol{\beta}}$ and set it to zero:

$$(2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} - 2\mathbf{X}'\mathbf{y}) + 2\lambda\hat{\boldsymbol{\beta}} = 0$$

Solving this equation would give the desired result.

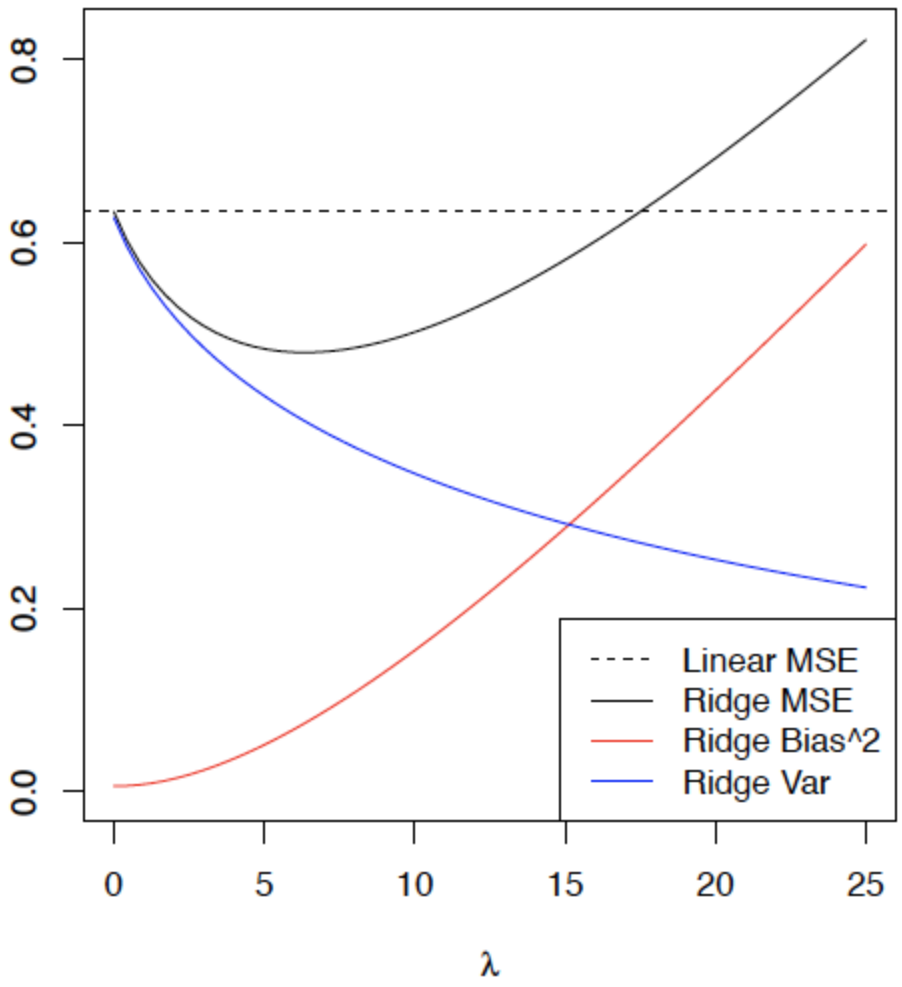
Properties of the ridge estimator

- The ridge estimator is biased:

$$\mathrm{E}(\hat{\boldsymbol{\beta}}_R) = (\mathbf{X}'\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

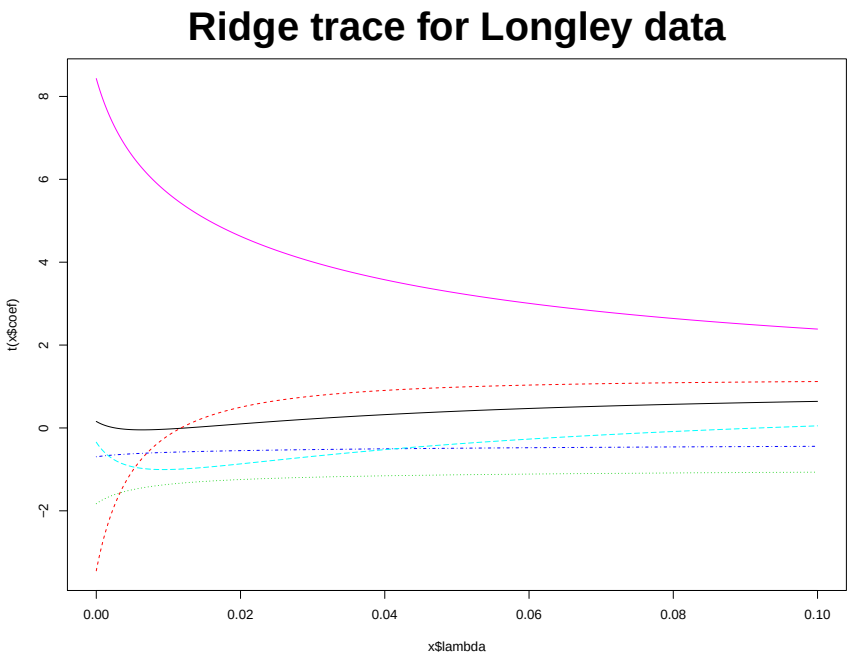
- As λ increases,
  - the bias increases, but
  - the variance decreases

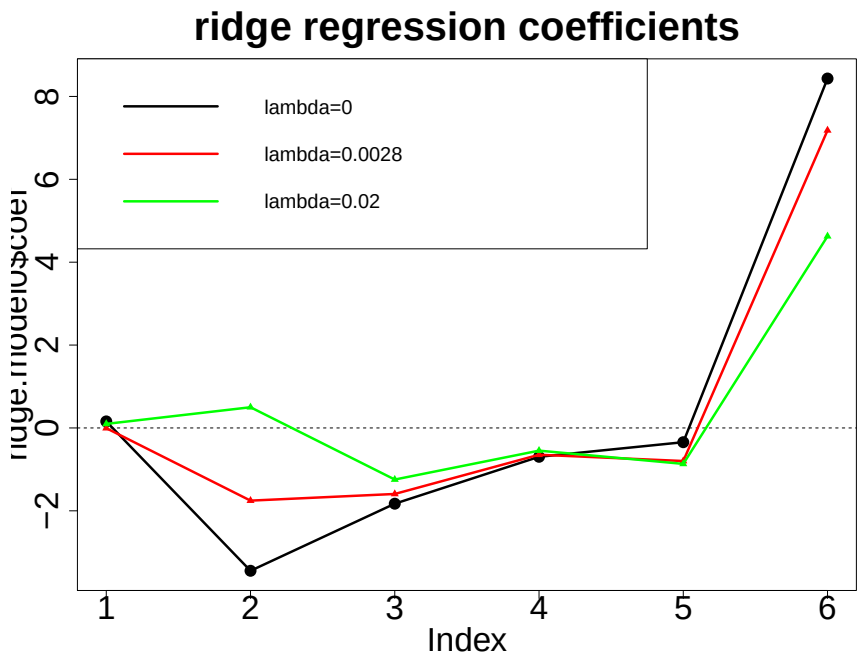
  Overall, ridge regression may still achieve a smaller MSE $(= \mathrm{Bias}^2 + \mathrm{Var})$.



Methods for setting the shrinkage parameter λ:

1. Ridge trace: Plot estimated coefficients in $\hat{\beta}_R$ against small positive values of λ, and select the smallest value of λ for which the parameter estimates have "stabilized".





Remark.

- Choosing λ by inspection of the ridge trace is a subjective procedure requiring judgment on the part of the analyst.
- The 'select' function can be used to select the value of λ:

  ## smallest value of GCV at 0.0028

- In your homework (Question 9.18), you are asked to try another empirical method for setting λ (Equation 9.8, page 313).

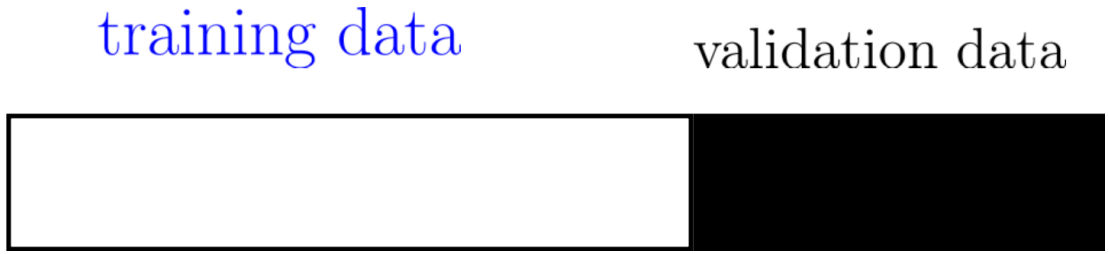
2. Model validation: Apply the corresponding regression models to an independent set of data and compare the resulting prediction errors

(1). **Confirmation runs**. Collect **new data** after the model has been fit to an existing data set

- If the existing model gives realistic predictions for the new data, users can have confidence in the validity of the model.
- There should be a good number of new points for testing the model and they should ideally be spaced out in predictor space.
- This validation method is most effective, but could be costly, or time consuming.

(2). **Data splitting**. A cheaper way of performing model validation by artificially/randomly partitioning the data into two parts:

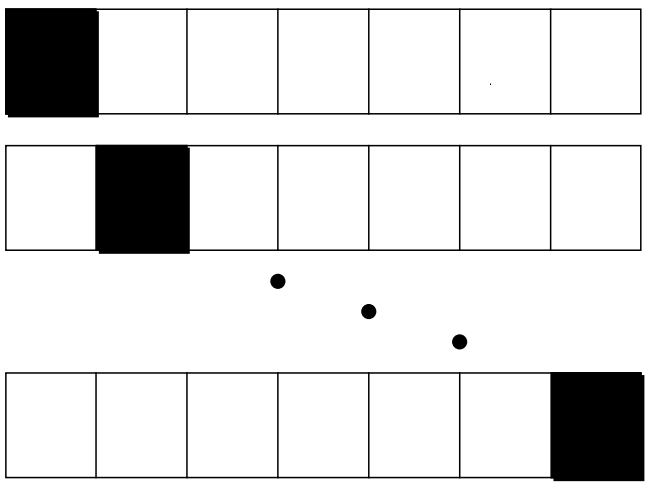
- estimation/training data: for fitting the model
- **prediction/validation data**: for evaluating the fitted model



Remark.

- Usually, the estimation set is larger than the validation set to ensure sufficient precision.
- If the data are collected over time (or over different locations), then time (or location) may be used as the basis of data splitting.
- A commonly-used data splitting scheme is called **cross validation**; see next slide for detail.

m-fold cross validation (CV):
Fix an integer m and partition the given data randomly into m equal-sized subsets (called folds).



When each fold has only 1 data point, it is called leave-one-out CV.

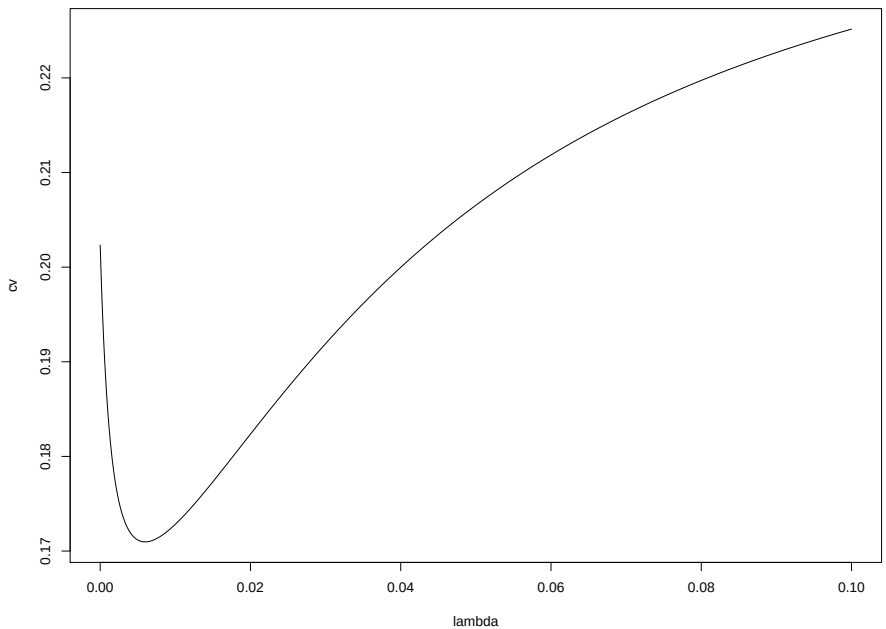
For each value of λ do the following:

- Of the m subsets, each one is retained as validation data once and the remaining $m-1$ are used as estimation data $\longrightarrow e_{\text{pred}}(i), 1 \le i \le m$
- Compute overall prediction error $e_{\text{CV}}(\lambda) = \sum_{i=1}^{m} e_{\text{pred}}(i)$ $\longrightarrow$ PRESS if $m = n$.

The optimal λ is the one minimizing the CV error: $\lambda^* \to \min_\lambda e_{\text{CV}}(\lambda)$

R code

```
#install.packages("parcor")
library(parcor)
ridge.object<-ridge.cv(X,y,
lambda=seq(0, 0.1, 0.0001),
scale=TRUE,
k=10,
plot.it=TRUE)
```



```
> ridge.object


$intercept

65.317

$coefficients

XGNP.deflator XGNP XUnemployed XArmed.Forces XPopulation XYear
-0.01935171 -1.60336592 -1.61615589 -0.66190834 -0.86949367 7.24516124

$lambda.opt

[1] 0.003325141
```

Recap: Ridge regression minimizes the usual regression criterion plus a penalty term on the squared L_2 norm of the coefficient vector. As such, it shrinks the coefficients towards zero. This introduces some bias, but can greatly reduce the variance, resulting in a better mean-squared error.

The amount of shrinkage is controlled by λ: Large λ means more shrinkage, and so we get different coefficient estimates for different values of λ. Cross validation is an effective way of choosing an appropriate value of λ.

It can be shown that ridge regression does not set coefficients exactly to zero unless $\lambda = \infty$ (in which case they're all zero). Hence, ridge regression cannot perform variable selection, and even though it performs well in terms of prediction accuracy, it does not offer a clear interpretation.

The LASSO

LASSO (Least Absolute Selection and Shrinkage Operator) differs from ridge regression only in terms of the norm used by the penalty term:

- **Ridge regression**:

$$\min_{\hat{\boldsymbol{\beta}}} \underbrace{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2}_{\text{fitting error}} + \lambda \underbrace{\|\hat{\boldsymbol{\beta}}\|^2}_{\text{penalty}} \longrightarrow \hat{\boldsymbol{\beta}}_R$$

  where the vector norm is the ℓ_2 norm: $\|\hat{\boldsymbol{\beta}}\| = \sqrt{\sum \hat{\beta}_j^2}$.

  We have pointed out that the ℓ_2 penalty only shrinks the coefficients but never forces them to be zero.

- **LASSO**:

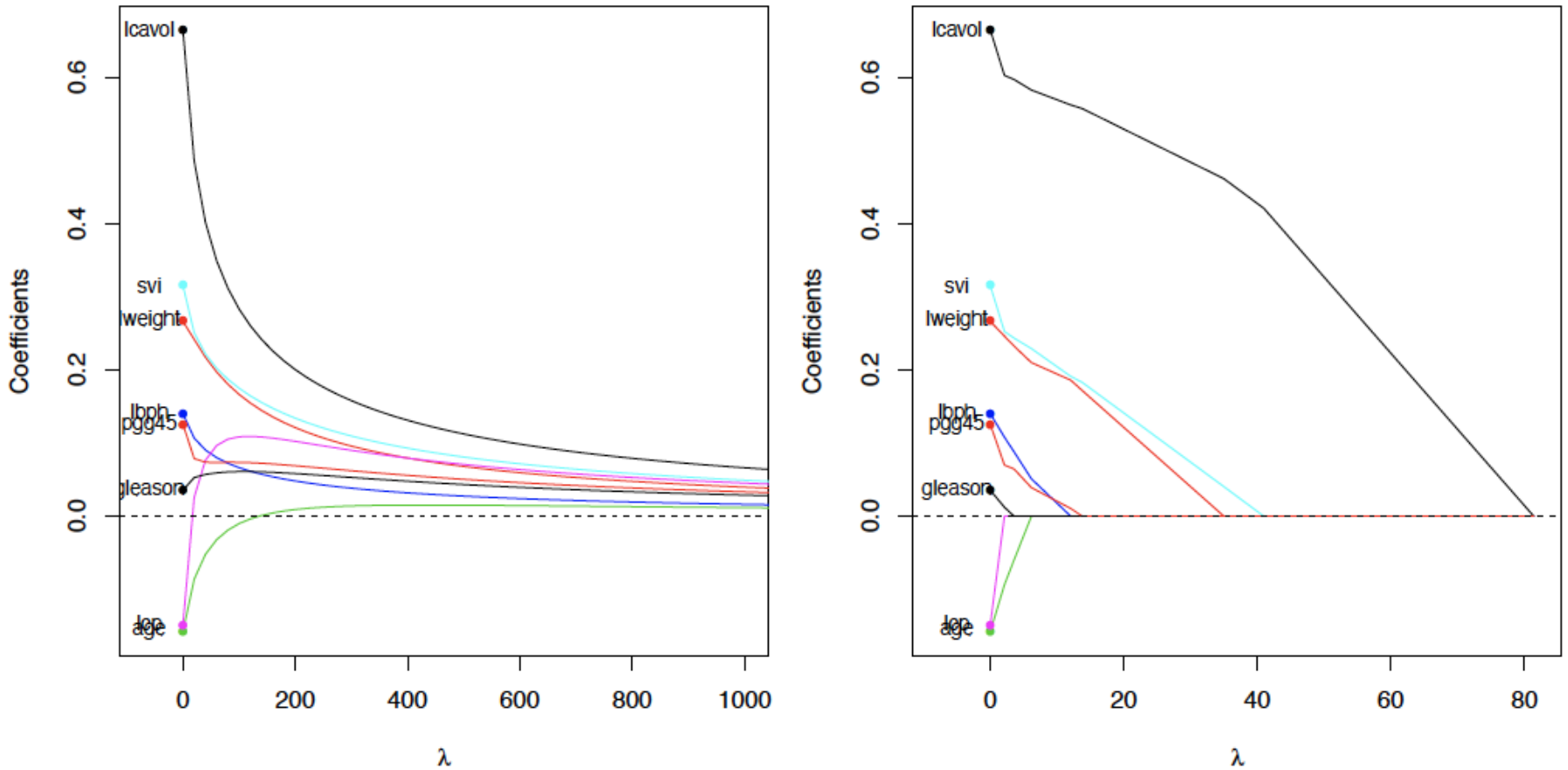
$$\min_{\hat{\boldsymbol{\beta}}} \underbrace{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2}_{\text{fitting error}} + \lambda \underbrace{\|\hat{\boldsymbol{\beta}}\|_1}_{\text{penalty}} \longrightarrow \hat{\boldsymbol{\beta}}_L$$

  where $\|\hat{\boldsymbol{\beta}}\|_1 = \sum |\hat{\beta}_j|$ is the ℓ_1 norm.

  The nature of the ℓ_1 penalty will cause some coefficients to be shrunken to zero exactly, and thus it is able to perform variable selection in the linear model: As λ increases, more coefficients are set to zero (less variables are selected), and among the nonzero coefficients, more shrinkage is employed.

In terms of prediction error (or mean squared error), lasso performs comparably to ridge regression, yet it has a big advantage w.r.t. interpretation.

Multicollinearity (and Model Validation)



(https://www.stat.cmu.edu/~ryantibs/datamining/lectures/17-modr2.pdf)

R code for lasso with the Longley data:

`https://www2.stat.duke.edu/courses/Fall16/sta721/slides/Lasso/lasso.pdf`

Summary

- **Multicollinearity**: Effects, detection, and treatment
- Ridge regression (not a model selection algorithm):

$$\min_{\hat{\beta}} \|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2 + \lambda\|\hat{\beta}\|^2 \quad \longrightarrow \hat{\beta}_R$$

- LASSO (can perform model selection):

$$\min_{\hat{\beta}} \|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2 + \lambda\|\hat{\beta}\|_1 \quad \longrightarrow \hat{\beta}_L$$

9.5.4 Principal-Component Regression (further learning)